Published online April 30, 2004

2464–2473 *Nucleic Acids Research, 2004, Vol. 32, No. 8*
DOI: 10.1093/nar/gkh566

Contact-based sequence alignment

Jens Kleinjung*, John Romein, Kuang Lin[1] and Jaap Heringa

Bioinformatics Unit, Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1081A, 1081HV Amsterdam, The Netherlands and [1]Division of Mathematical Biology, National Institute for Medical Research, Mill Hill, London NW7 1AA, UK

Received March 15, 2004; Revised and Accepted April 5, 2004

ABSTRACT

This paper introduces the novel method of contact-based protein sequence alignment, where structural information in the form of contact mutation probabilities is incorporated into an alignment routine using contact-mutation matrices (CAO: Contact Accepted mutatiOn). The contact-based alignment routine optimizes the score of matched contacts, which involves four (two per contact) instead of two residues per match in pairwise alignments. The first contact refers to a real side-chain contact in a template sequence with known structure, and the second contact is the equivalent putative contact of a homologous query sequence with unknown structure. An algorithm has been devised to perform a pairwise sequence alignment based on contact information. The contact scores were combined with PAM-type (Point Accepted Mutation) substitution scores after parameterization of gap penalties and score weights by means of a genetic algorithm. We show that owing to the structural information contained in the CAO matrices, significantly improved alignments of distantly related sequences can be obtained. This has allowed us to annotate eight putative *Drosophila* IGF sequences. Contact-based sequence alignment should therefore prove useful in comparative modelling and fold recognition.

INTRODUCTION

Protein sequence alignments reveal information about the relation between homologous sequences. Closely related sequences yield mostly unambiguous alignments, but distantly related sequences with sequence identities between 15 and 30% (the so-called twilight zone) pose a difficult problem to sequence alignment routines, because the signal of similarity is heavily disturbed by 'noise' from mutations (1). Protein structure alignments serve as standard of truth for judging evolutionary relatedness, since structures are much better conserved than sequences (2). Although the structure of most sequences is unknown, it can be inferred in many cases by comparative modelling and fold recognition. Comparative modelling delivers useful models that can attain a quality comparable to experimentally derived structures. However, the quality is crucially dependent on the accuracy of the underlying sequence alignment (3), which is exponentially decreasing with increasing evolutionary distance of the related sequences (4).

In conventional sequence alignment (5–17), structural information available from the template is ignored, because alignment programs use PAM-type (Point Accepted Mutation) substitution matrices (18), which incorporate only sequence information. For the purpose of improving sequence alignment accuracy, we introduce here an alignment method that uses structural information from side-chain contacts. Contacts describe constraints between residues and particularly long-range contacts can help defining the relative arrangement of sequence elements. Alignment scores are provided by the CAO substitution matrices (Contact Accepted mutatiOn), which are derived from an evolutionary Markov model of side-chain contact substitution (19). The CAO matrices describe the mutation probability of side-chain contacts within a protein and therefore they combine sequence and structure information in a single scoring scheme. CAO scores often reflect evolutionary relatedness better than PAM scores. We introduce a contact-based sequence alignment algorithm, which transfers information from a template protein structure to the alignment process by means of CAO scores. The predictive power of the method is illustrated by the annotation of eight putative *Drosophila* IGF sequences through alignment to a human IGF-I sequence. The annotation is supported by a comparative model based on the contact-based alignment and independently generated *de novo* structural models.

MATERIALS AND METHODS

Programs were written in the 'ANSI C' programming language and compiled with the GNU or Intel compiler on Linux 2.4.

The CAO contact matrix

The CAO substitution matrix is based on a Markov model of protein side-chain contact evolution (19). CAO scores are an intermediate between the purely sequence-based PAM scores and the purely structure-based 'Root Mean Square Deviation' (RMSD) values. The CAO matrix comprises 400×400 contact substitution scores, where rows and columns consist of the 20×20 possible combinations of residue contacts, and each matrix cell contains a score for the evolutionary transition (mutation) from a contact denoted by the row axis

*To whom correspondence should be addressed. Tel: +31 20 4447783; Fax: +31 20 4447653; Email: jkleinj@cs.vu.nl

to a contact denoted by the column axis. Thus, knowing the side-chain contacts within a template structure, an alignment of the template sequence with a query sequence can be scored by summing up the CAO substitution matrix values of all contacts.

The contact-based alignment algorithm

Structural information of the template is passed to the alignment program in the form of a side-chain contact map. A contact is defined if the distance between two side-chain spheres is smaller than the diameter of a solvent molecule (19), which is 2.8 Å for water. A sequence alignment method using the CAO matrix considers contact pairs consisting of four residues: two residues that are known to contact in the template sequence are aligned to two contact residues of the query sequence. For example, if the template residues A and B are in contact (A_B), and aligned to query residues X and Y, the CAO score of A_B:X_Y can be assigned ('_' indicates a contact, ':' indicates an alignment).

Needleman-Wunsch-type dynamic programming is not suited for using CAO contact scores directly, because the alignment of the second matched pair (B:Y) needs to be known to correctly score the first matched pair (A:X) and vice versa. This violates the independence assumption, i.e. the prerequisite that the score of each matrix cell should be independent of any other cell.

Therefore, we devised an alignment algorithm that incorporates CAO contact scores by pre-processing into a dynamic programming (DP) matrix, such that DP can be performed on this matrix after pre-processing. The algorithm works as follows: the structure of the query sequence (containing residues X and Y) is unknown and therefore the position of the putative homologous contact X_Y in the query is unknown. Thus, the query has to be probed for potential positions of this contact by testing all possible realizations on the query sequence, assuming that the query residues are in contact as well (X_Y) (Fig. 1). When this scheme is applied to a DP matrix, all contacts of the template are slid over the query sequence (Fig. 2). Each possible (and hypothetical) realization of a contact is probed with CAO scores: likely contact mutations yield positive scores, unlikely contact mutations yield negative scores. At each position of the sliding contact, the CAO score of the contact match is added to each of the two DP matrix cells involved. Most residues form multiple contacts, and their scores are summed up in the matrix. After scoring all contacts (pre-processing), the optimal alignment is found by forward score addition and back-tracing as in the standard DP alignment algorithms. Routines for local and global alignment have been implemented. To compensate for potentially missing local contact information in the template, the DP matrix is complemented with PAM-type substitution matrix scores.

The algorithm can be summarized as follows: (i) Slide each contact of the template over the query sequence and, at each position, add the (weighted) CAO score of the four involved residues to the two corresponding matrix cells. (ii) Fill the alignment matrix with PAM-type substitution scores. (iii) Find the optimal path through the matrix by forward score addition and back-tracing.

A complication is the use of float values in the CAO matrix, which can, without precaution, lead to rounding errors and failure of the DP routine during the traceback phase. This has been solved by incorporating a small allowance for score deviations caused by rounding errors.

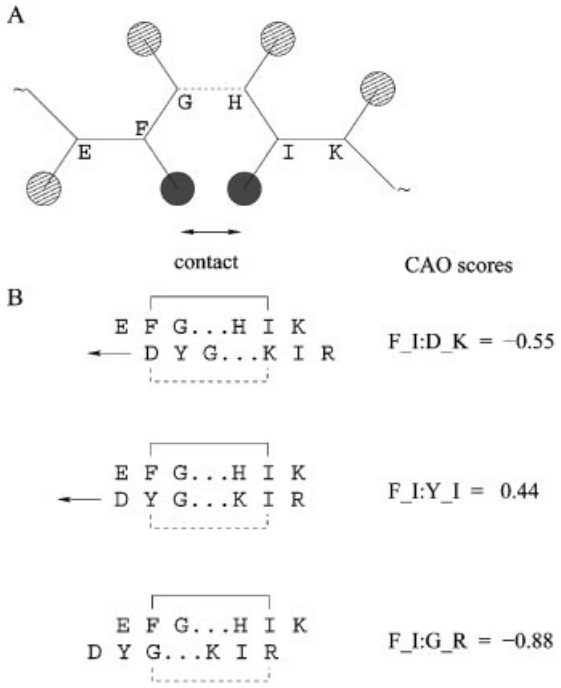

Figure 1. Principle of contact information transfer from structure to sequence. (**A**) Scheme of template protein structure with contact between side-chains of residues F and I. (**B**) The same contact is marked on the template sequence (EFG...HIK) by the solid bracket between F and I. Alignment of the homologous query sequence (DYG...KIR) implies structural equivalence of the aligned residue pairs. Therefore, a contact can be tentatively assumed to exist in the query between D and K (dashed bracket). The CAO matrix yields a score for the contact match F_I:D_K. The putative contact (dashed bracket) is probed at all possible positions of the query sequence, which is illustrated here by sliding the query sequence along the template sequence as indicated by the arrows.

Benchmarking

The contact-based alignment was benchmarked using the Homstrad database (20). Only families containing two single-chain sequences were considered (624 alignments in total), since multiple alignments could change the pairwise relationships: the result of aligning two sequences in an optimal pairwise alignment or in a multiple alignment is not necessarily identical. Of the 5304 PDB structures used for the design of the CAO matrices, 355 (6.7%) were contained in the Homstrad benchmark set. The fraction of identical alignment positions when comparing the CAO alignment with the reference Homstrad alignments served as quality measure (or fitness function) and was reported as 'fraction correct pairs'. The T-Coffee alignment program (version 1.37) was used for comparison (16). T-Coffee is invariably the top performing algorithm in recent multiple sequence alignment benchmarks.

Parameterization

To scale the CAO scores relative to PAM-type scores from Blosum62 (21) or PAM120 (18), four parameters were optimized by running the CAO alignment algorithm using a

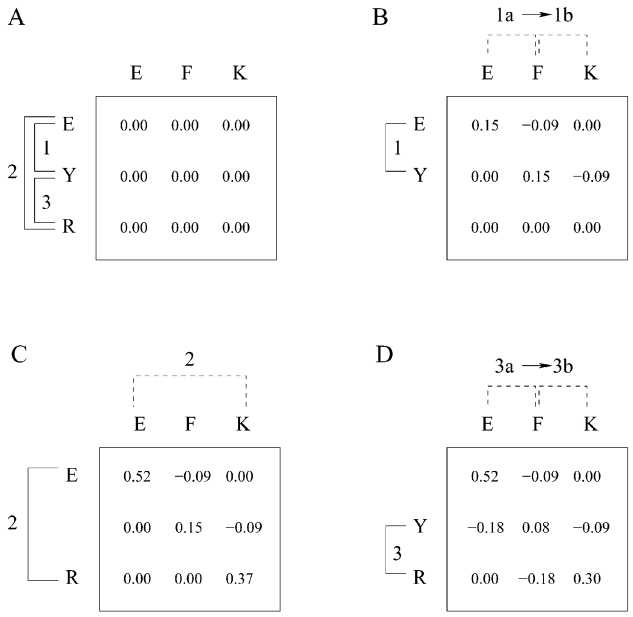

Figure 2. Illustration of the contact-based alignment by dynamic programming. Newly assigned scores are highlighted in bold font. (**A**) The template sequence EYR contains three contacts: E_Y (1), E_R (2) and Y_R (3). The query sequence is EFK. (**B**) Contact (1) is slid over the query sequence, realising positions (1a) and (1b). In position (1a), the CAO score for E_Y:E_F yields score 0.15 for the two matrix cells EE and YF on the main diagonal. Position (1b) yields –0.09 for E_Y:F_K. (**C**) Contact (2) has only one possible realization in this example. The score of E_R:E_K is 0.37 as can be seen in cell RK, whereas the score of cell EE is 0.37 + 0.15 = 0.52. (**D**) Contact (3) is comparable to contact (1), with scores –0.18 for Y_R:E_F and –0.07 for Y_R:F_K.

genetic algorithm (GA): a CAO matrix constant c, the relative weight w of CAO versus PAM-type scores, gap-opening penalty p and gap extension penalty q. Thus, CAO and PAM scores are combined to

$s_{CAO}(i,k) = s_{PAM}(i,k) + ws_{CAO}(i,j,k,l) + c$ and $s_{CAO}(j,l) = s_{PAM}(j,l) + ws_{CAO}(i,j,k,l) + c$

while gap penalties p,q influence the final path through the matrix. The GA was performed with a population size of 100 individuals (genomes) over eight generations, where the 20 fittest individuals were selected for mating. Mating was performed on randomly chosen parent pairings (of the 20 fittest individuals) and the child inherited parameters from either parent at random. An individual's genome was composed of four parameter values (genes). The starting population was random with target ranges (step width in parentheses): CAO matrix constant c [0–0.9 (0.1)], relative weight w [0–0.45 (0.05)], gap-opening penalty p [5–14 (1)] and gap-extension penalty q [0–4.5 (0.5)]. Each GA run reached convergence within eight generations and was repeated three times. To prove the optimized parameters' independence of the particular sequence set, the GA parameterization was additionally jack-knifed by splitting the database into two parts (first jack-knifing) and five parts (second jack-knifing). The variation of the parameters is given as standard deviation in Table 1.

Drosophila IGFs

Drosophila protein sequences were downloaded from the Ensembl site (ftp://ftp.ensembl.org/pub/current_fly/data/fasta/pep/). The first stage of the identification was achieved by pattern matching using the regular expression */C\w{11,14}C\w{4,}C\w{12,15}CC/* that flexibly matches the canonical inter-chain disulfide bridges of insulin-like proteins. In the second stage, the programs Psi_Blast (22) and Quest were used (23) with relaxed parameters (E-value < 10), since standard sequence searching methods failed to detect the (low) sequence similarity. The protein access codes of the identified sequences are AAF47991 (DIGF-1), AAF47993 (DIGF-2), AAF48005 (DIGF-3), AAF48006 (DIGF-4), AAF48007 (DIGF-5), AAF48078 (DIGF-6), AAF50020 (DIGF-7) and AAF51015 (DIGF-8). Most of the DIGFs are surely expressed as the cDNAs of all but DIGF-3 and DIGF-7 are known. Sequence alignments were produced using the contact-based sequence alignment algorithm with a CAO weight of 0.2. Comparative modelling was performed using the DeepView program (24). The DIGF-3 query sequence was modelled onto the crystal structure of hIGF-I (PDB databank entry 1GZR) (25). Ten *de novo* structural models for the DIGF-3 sequence were predicted by the ROBETTA structure prediction server (26).

RESULTS

Benchmarking

The improvement of the alignment quality in terms of structurally correct sequence relation was assessed using

Table 1. GA-optimized parameters and total scores of parameterization runs over 624 pairwise alignments of the Homstrad database for the contact-based alignment routine (CB) and Needleman-Wunsch dynamic programming (NW)

	Matrix[a]	p[b]	q[b]	w[c]	c[d]	Max. score[e]	f.c.[f] (%)
CB	B + C	14.0 ± 0.8	1.0 ± 0.5	0.1 ± 0.05	2.8 ± 0.5	510.6	81.8
CB	P + C	14.0 ± 1.0	1.0 ± 0.5	0.2 ± 0.1	2.4 ± 0.5	502.5	80.5
NW	B	11.0 ± 1.2	1.5 ± 0.5	–	–	492.0	78.8
NW	P	10.0 ± 0.8	1.5 ± 0.5	–	–	465.5	74.6

Alignments were performed using CAO (C) and Blosum62 (B) or PAM120 (P) matrices.
[a]Substitution matrices; B: Blosum62, C: CAO120, P: PAM120.
[b]Gap opening parameter p and gap extension parameter q.
[c]Relative weight w of CAO versus Blosum62 or PAM120 scores.
[d]Matrix constant c for CAO scores.
[e]Maximal sum of 'fraction correct pairs' scores over the benchmark set.
[f]Average of 'fraction correct pairs' scores.

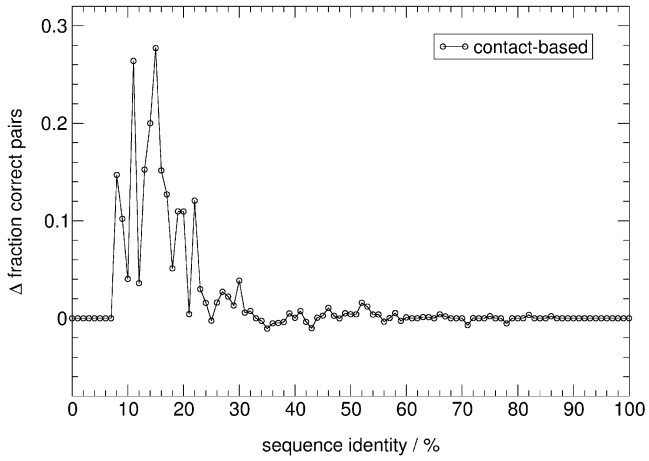

Figure 3. Benchmark of alignment routines. Difference Δ between the 'fraction correct pairs' as produced by contact-based alignment and by Needleman-Wunsch alignment. Improvement of alignment quality is apparent at low sequence identity, where contacts add valuable information to the sequence alignment.

pairwise alignments from the Homstrad database (20) as standard of truth. The agreement between the contact-based alignments with the Homstrad alignments was calculated as the fraction of identical alignment positions, given as 'fraction correct pairs'. Optimized parameters resulting from GA runs contact-based alignment are shown in Table 1. Each CAO score receives a weight factor *w* of 0.1 (in combination with Blosum62) to 0.2 (in combination with PAM120). Thus, CAO scores are one-tenth to one-fifth as important as PAM scores, but taking into account that each residue has about seven contacts, the overall contribution per residue is similar. The Blosum62 matrix performs better (81.8%) than the PAM120 matrix (80.5%). It is reasonable to use the CAO matrix with positive scores in combination with the Blosum62 matrix. The difference to the 'fraction correct pairs' from Needleman-Wunsch alignment is plotted as a function of sequence identity in Figure 3. Alignment quality is improved in the region of low sequence identity.

We compared our method to the multiple alignment program T-Coffee (16), which also yielded significant improvements over Needleman-Wunsch alignment, although for a different reason. The contact-based alignment routine adds structural information to the alignment, while T-Coffee combines local and global sequence information to optimize the alignment. Benchmark results are summarized in Table 2. The contact-based program performs well on the benchmark set and improves 64% of the sequences when compared to Needleman-Wunsch alignment, and with 16% more alignments than T-Coffee. The computational speed of the contact-based alignment is 1.6-fold slower than Needleman-Wunsch alignment, measured for a benchmark run over 624 sequences, but twice as fast as T-Coffee on the same sequence set. It is interesting to analyse the correlation between the contact-based alignments and T-Coffee alignments. The change in alignment quality of both methods was monitored by counting the number of correlated changes: 276 alignments (43%) were improved, 76 (12%) were worsened and 78 (12%) were unchanged by both methods. About one-third of all alignments [203 (33%)] showed no correlation. Thus, a combination of the two methods should further improve the quality of alignments with low sequence identity.

Table 2. Benchmark of alignment routines: Needleman-Wunsch, contact-based and T-Coffee

	Needleman-Wunsch	Contact-based	T-Coffee
Fraction correct (%)	78.8	81.8	80.8
Improved	–	400 (64%)	299 (48%)
Unchanged	–	116 (18%)	115 (18%)
Worsened	–	108 (18%)	210 (34%)
cpu time (s)	12.6	19.6	40.3

Numbers are given for 624 pairwise alignments of the Homstrad database. Performance comparison is made against Needleman-Wunsch alignments.

An example for the improvement by contact information is shown in Figure 4. The C-terminal domain of the protein chondroitinase (Homstrad family 'Lyase_8_C') forms a β-sandwich, containing nine strands in *Flavobacterium heparinum* (top sequence) and eight strands in *Streptococcus pneumoniae*. Repetitive secondary structure elements often exhibit a spurious sequence similarity, resulting in register-shifted alignments as shown in the T-Coffee example. In contrast, the contact-based alignment keeps the alignment in register except for the C-terminal short strand, and thus dramatically improves the alignment. The predominant $1 \rightarrow 3$ side-chain contacts in β-sheets each correlate four residues and register shifts are less likely to yield high scores.

Evaluation of contacts

By sliding contacts over the query sequence, we implicitly make the assumption that the distance between the contacting residues has been conserved. This assumption keeps the contact-based alignment algorithm computationally efficient and it holds for the majority of contacts, but obviously not for locations with insertion/deletion (indel) events, which disrupt the original local contact pattern. Contacts with changed residue distance add noise to the DP matrix by scoring non-homologous contact pairs, in most cases with negative scores. Figure 5A,B illustrates the conservation of sequence and structure information. Whereas the fraction of correctly aligned residues decreases toward zero with decreasing sequence identity (Fig. 5A), the fraction of conserved contacts (including the contact length) approaches 30% (Fig. 5B). This is an illustration of the commonly known fact that 'structure is more conserved than sequence' (2), and it supports the above approximation of conserved contact length. Figure 6A shows the distribution of contact distances averaged over the benchmark set of proteins from the Homstrad database. Many of the contacts are short ranged (about 24%), within a distance of 1–5 residues. The highest peaks (distances 1 and 3) originate from contacts within helices. Strands account for most of the contacts at a distance of two residues. The long-range contacts are quite evenly spread out over a large range of distances. Long-range contacts can be important in defining the correct relative arrangement of distant sequence elements, which is hard to achieve with classical sequence alignment. The distribution of contact numbers is given in Figure 6B. Most residues form about seven contacts, but the distribution

```
Homstrad
.PKVLANTNQLQAVYHQQLDMVQAIFYTAGKLSVAGIEIETDKPCAVLIKHI
ESSLIENNETLQSVYDAKQGVWGIVKYDDSVSTISNQFQVL.KRGVYTIRKE
NGKQVIWAADPLQKEKTA...VLSIRDLKTGKTNRVKIDFPQQEFAGATVEL
GDEYKIAYYNPETQESAPDQEVFKKQH.........................

T−Coffee (0.18)
.PKVLANTNQLQAVYHQQLDMVQAIFYTAGKLSVAGIEIETDKPCAVLIKHI
ESSLIENNETLQSVYD.............AKQGVWGIVKYDDSVSTI.....
NGKQVIWAADPLQKEKTAVLSIRDLKTGKTNRVKIDFPQQEFAGATVEL.....
........SNQFQVLKRGVYTIR..KEGDEYKIAYYNPETQESAPDQEVFKKQH

Contact−based (0.69)
.PKVLANTNQLQAVYHQQLDMVQAIFYTAGKLSVAGIEIETDKPCAVLIKHI
ESSLIENNETLQSVYDAKQGVWGIVKYD.DSVSTISNQFQVLKRGVYTIRKE
NGKQVIWAADPLQKEKTAVLSIRDLKTGKTNRVKIDFPQQEFAGATVEL
GDEYKIAYYNPETQE.....SAPDQEVFKKQH.................
```

Figure 4. Sequence alignments of the C-terminal domain of the protein chondroitinase (Homstrad family 'Lyase_8_C'). Numbers in parentheses show the 'fraction correct pairs'. The top alignment represents the reference alignment derived from structural super-positioning. Secondary structure elements of the β-sheet sandwich structure are indicated by horizontal lines. The T-Coffee alignment reveals a register shift in the matched secondary structure elements, which is prevented in the contact-based alignment except for the short C-terminal strand of the bottom sequence.

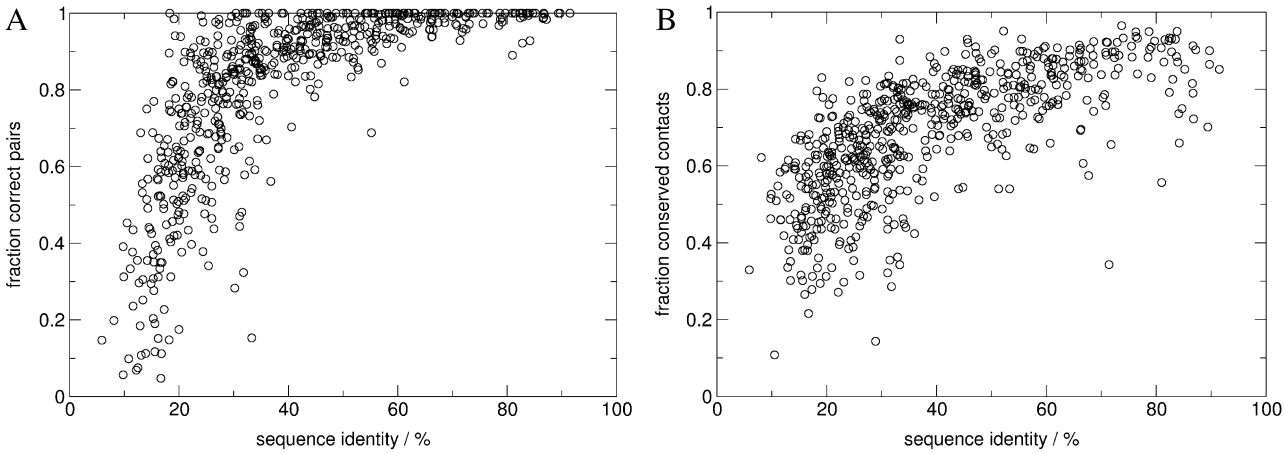

Figure 5. Comparison between sequence and structure information. (**A**) Fraction of correctly aligned residue pairs as a function of sequence identity. Homstrad families containing two sequences (624 families) were aligned by Needleman-Wunsch dynamic programming using the Blosum62 matrix and gap penalties 14,1 (opening and extension) and compared to the reference Homstrad alignments. Alignment quality decreases exponentially with decreasing sequence identity. (**B**) The same set of families as in panel (A) was used to compare the contacts within homologous protein structures. A contact is conserved when it is found at the same relative position in both structures and consequently at aligned positions in the reference sequence alignment (which is derived from structural super-positioning) as well. Even at very low sequence identity (10–20%), about 30% of contacts are conserved.

is quite broad, leading to considerably different contributions of residues to the DP matrix.

In this context, it is instructive to plot the dependence of contact conservation upon contact distance, as given in Figure 7A. Surprisingly, the contact conservation is nearly invariant to the contact distance, with the exception of short-ranged helical contacts, which give rise to the peaks at distances 1 and 3 (compare to Fig. 6). However, the level of

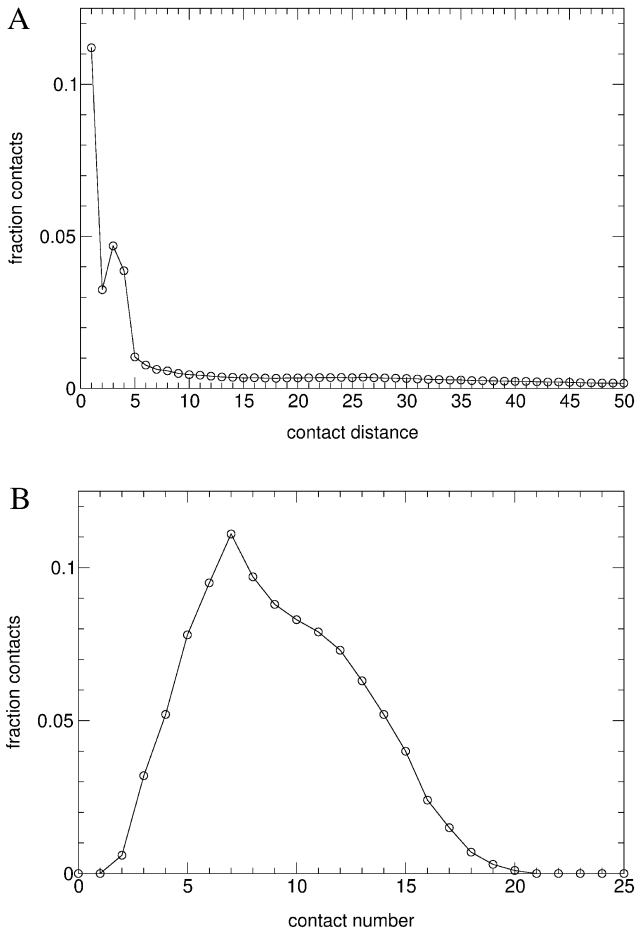

Figure 6. Average distribution of contact distance and contact number. (**A**) The contact distance is defined as the difference of residue numbers between two contacting residues. The distribution is dominated by helices, whose 1 → 2 (distance = 1) and 1 → 4 (distance = 3) side-chain contacts are most frequent. Although most contacts are short-ranged, long-range contacts are generally present and provide valuable information about the packing of secondary structure elements. (**B**) The contact number is defined as the number of contacts a single residue forms with other residues.

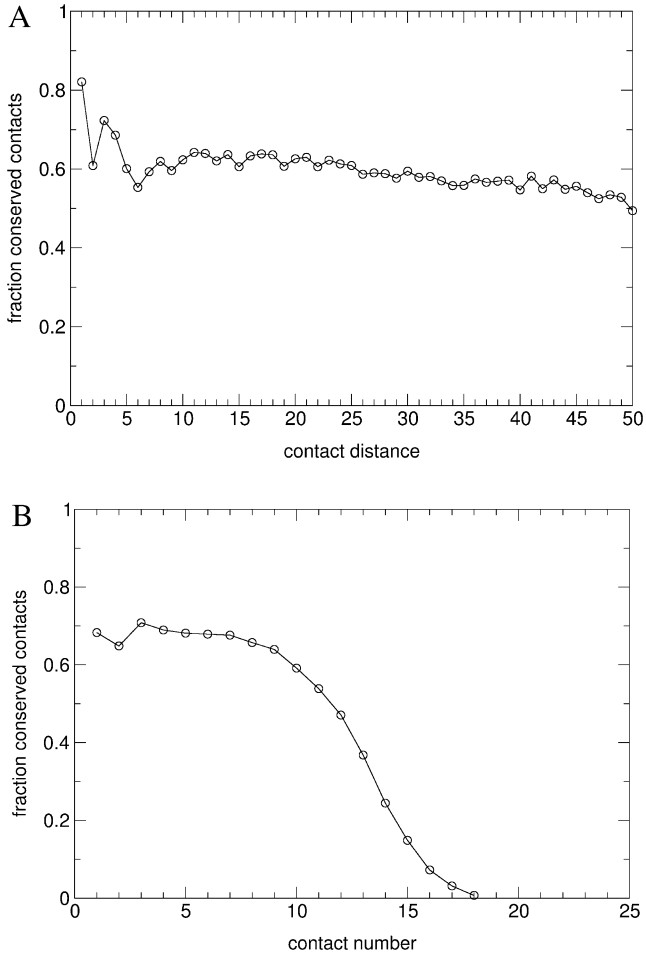

Figure 7. Contact conservation. (**A**) Fraction of conserved contacts as a function of contact distance. Contact conservation is marginally influenced by the contact distance. The short-range pattern originates from intra-helical contacts, which exhibit a better conservation than other contacts. (**B**) Fraction of conserved contacts as a function of contact number. Conservation is constant from 1 to 8 contacts per residue, after which it decreases rapidly. Residues forming many contacts seem to accept mutations more than those with few contacts.

conservation is very dependent upon the number of contacts that each residue forms. The curve in Figure 7A reveals a good level of conservation for residues with 1–8 contacts, and a steep decrease in conservation at higher contact numbers. A plausible explanation is that single contacts can be more easily replaced without losing essential interactions, because each single contact is less important.

Drosophila IGFs

Contact-based sequence alignment proved instrumental in our annotation of insulin-like *Drosophila* genome sequences. Insulin-like growth factors (IGFs) and insulin are hormones belonging to the insulin protein family. IGFs are essential growth factors in development, whereas insulin acts as a regulator of the blood glucose level (27). The main structural difference between IGF and insulin is the post-translational tryptic cleavage of the C-peptide between A- and B-chain in insulin, while IGF remains a single-chain protein due to the lack of tryptic cleavage sites. The insulin protein family occurs predominantly in vertebrates, although insulin-like proteins are known in lower organisms, for example molluscan MIP (28), locust LIRP (29), bombyxin in the silkworm *Bombyx mori* (30), insulin-like peptides in the nematode *Caenorhabditis elegans* (31) and DILPs in the fruit fly *Drosophila* (32). The reported DILP proteins possess tryptic cleavage sites in the C-peptide and therefore they are considered as insulin-type proteins.

We report here the identification of eight putative IGF-type sequences (DIGFs) in the genome of *Drosophila* (33) that have no tryptic cleavage sites. A multiple alignment of the DIGF sequences is shown in Figure 8. The DIGFs exhibit a cysteine pattern that resembles closely the characteristic pattern of the insulin family. However, an excess amount of cysteines and the low sequence identity between human IGF (hIGF-I) and DIGFs render the matching of cysteines ambiguous when using sequence-only alignment.

```
DIGF-1        MPIDRQFYDEFSETTPIGVRDQFSPVSNTKAVAMHPERYASPNPLIPLVIAGALLFSMLS
DIGF-8        ----------------------------------------------------MLLWILFAA
DIGF-2        ------------------------------------------------MESISSMIYLVA
DIGF-3        ----------------------------------MSGAAWMSLAFVACLLAASVDGN
DIGF-4        ----------------------------------------MRAGVACLLVLLGICGAA
DIGF-6        ----------------------------------MSQFSTVAAFLLLGLVVILGGHV
DIGF-7        ------------------------------------------------------------
DIGF-5        ------------------------------------MHNRCGSIWLLAAVLLLLAL

DIGF-1        QVSGYSGRIPPDADNPGKCMYRG--------DVLELGVNNGIAP------C-QRLTCNKD
DIGF-8        VQSAEWEDIHYDPNHPGKCTINP----------GLVLNPGVSIKDPTHEC-RKILCGLN
DIGF-2        MMSLIIGGSQAIPYRPSAYLYNQ----QYCMDTLTGRQLYIGEVFTREDQC-VRIQCLET
DIGF-3        GFSTQYRGHTQHPSLAEHCLYEE--------LDLAVPLNGYVLPSGQQGYC-IRLECTDD
DIGF-4        RADLTYRGNAVHPDYPGQCYYEE--------LNQAIPKKQSYKPINREGYC-QSIYCRPD
DIGF-6        GQAAVAKVKLNNSSHPGKCVLDT--------NTILSPGETGLAP---DLPC-VRAECHAD
DIGF-7        -------MSSQDRAHPGKCFDK--------LTRKALLPDKEYKP---KGIC-AAMTCSLE
DIGF-5        LLPQALLPTVDAASEPVCSYRNSEDETIFLKYLPLLRRGQDYVDFGKDGKCLKRAICTDT
                             *                                  *     *

DIGF-1        G-SILIEGCGKLRIE--NCNRGERISPGEPFPECCKLRYKCKQIGAAPYYIERNTAEKV-
DIGF-8        G-RVVYHSCGVSILS-PPCRYGDYINPDLPYPDCCSRTLLCN------------------
DIGF-2        L-QLWEDSCQVPKLTQGNCTPVPSTNPHAEYPRCCPLYECKSYESNSGGTLEQTNIYDHY
DIGF-3        Y-LLLIRHCDKQPWPRPGCHLSPN-DYDFKFPECCPQLECSDEY----------------
DIGF-4        Y-VLEISYCGRHNLVPTEKCRIAS-DMRRTFPECCPKLVCQESESNYI------------
DIGF-6        G-LVTFKTCDAVAPP-PGCKQRDFVNINREFPACCERKYNCDKHI---------------
DIGF-7        ALEISIETCPY--VEAPGCEELPS-DPNWRFPKCCPQFKCVDFKTGKDFIVSL-------
DIGF-5        F-KTIVEDCGQQKVT---CGNKDRFT--GVFPACC--LKCP-------------------
                     *         *              **    *

DIGF-1        ----------------------------------
DIGF-8        ----------------------------------
DIGF-2        GTLRSSHLTEMIVIDGRTPPRGEIHTASARKYQV
DIGF-3        ----------------------------------
DIGF-4        ----------------------------------
DIGF-6        ----------------------------------
DIGF-7        ----------------------------------
DIGF-5        ----------------------------------
```

Figure 8. Multiple sequence alignment of putative DIGFs. Positions with conserved cysteines are indicated with a star.

Contact-based sequence alignment between the structure of hIGF-I and DIGFs yields an overall consistent picture (Fig. 9A). The multiple alignment between the hIGF-I sequence and some DIGFs was derived from pairwise contact-based alignments, and it illustrates the corresponding cysteine patterns (filled boxes); the connectivity is shown in Figure 9B. The canonical inter-chain disulfide bridges C^{A7}–C^{B7} and C^{A20}–C^{B19} are conserved in DIGFs, but not the intra-chain disulfide bridge C^{A6}–C^{A11}.

We have generated a structural model of DIGF by means of comparative modelling, based on the alignment in Figure 9A and the crystal structure of hIGF-I (25). The model shows that the spatial proximity of cysteines C^{A19} and C^{B24} allows for the formation of an additional inter-chain disulfide bridge. A slightly different pairing C^{A19}–C^{B19} and C^{A20}–C^{B24} has been proposed for insulin-like proteins of *C. elegans* (31). Both bonding patterns seem to be viable in principle when judged on the basis of our comparative model; however, the pattern proposed in Figure 9B appears more favourable.

To further investigate our hypothesis about the insulin-like nature of the DIGFs, we submitted the DIGF-3 sequence to the ROBETTA structure prediction server that produces *de novo* structural models for submitted sequences (26). Analysis of the 10 predicted structures reveals predominant topological features of A- and B-chain. The superposition of a selected model (Fig. 10, yellow) with the hIGF-I chains (blue) illustrates the close resemblance. The B-chain is predicted to be α-helical in agreement with insulin-like B-chains. The A-chain tends to adopt the helix-coil-helix motif that is typical for the A-chain of insulin-like proteins.

A

```
                   B−chain                      ─C−peptide─
           7                         19    24
IGF-IA     .ETLCGAELVD.A..L...........QFVCGDRGFYFNKPTGYGSSSPQT
DIGF-3     LAEHCLYEELDLAVPLNGYVLPSGQ..QGYCIRLECTDDYLLLIRHCDKQP
DIGF-4     YPGQCYYEELNQA..IPKKQSYKPINREGYCQSIYCRPDYVLEISYCGRHN
DIGF-7     HPGKCFDKLTRKA..LLPDKEYKP...KGICAAMTCSLEALEISIETCPYV

                   A−chain
            6 7   11        1920
IGF-IA     GIVDECCFRSCDL.RRLEMYCAPLKP........
DIGF-3     WPRPGCHLSPNDYDFKFPECCPQLECSDEY....
DIGF-4     LVPTEKCRIASDMRRTFPECCPKLVCQESESNYI
DIGF-7     EAPGCEELPSDPN.WRFPKCCPQFKCVDFKTGKD
```

B

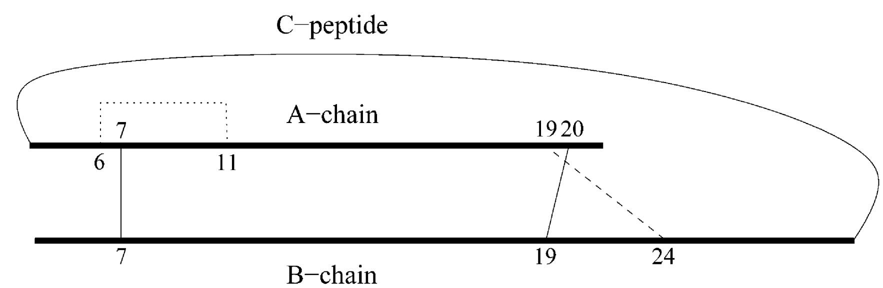

Figure 9. Primary structure of putative DIGFs. (**A**) Multiple sequence alignment of the hIGF-I sequence with DIGFs. The alignment was constructed from pairwise contact-based alignments using the crystal structure of hIGF-I as template (25). Numbering is according to the hIGF-I sequence. The N-terminal residues 'GP' (B-chain) and residues 'RRAP' (36–39, C-peptide) are not resolved in the structure. (**B**) Schematic illustration of chain structure and disulfide bridges in IGFs. The C-peptide connects the C-terminus of the B-chain with the N-terminus of the A-chain. Inter-chain bridges C^{A7}–C^{B7} and C^{A20}–C^{B19} are present in hIGF-I and DIGFs. The intra-chain disulfide bridge C^{A6}–C^{A11} is absent in DIGFs (dotted line). However, an alternative inter-chain disulfide bridge C^{A19}–C^{B24} might be formed in DIGFs (dashed line).

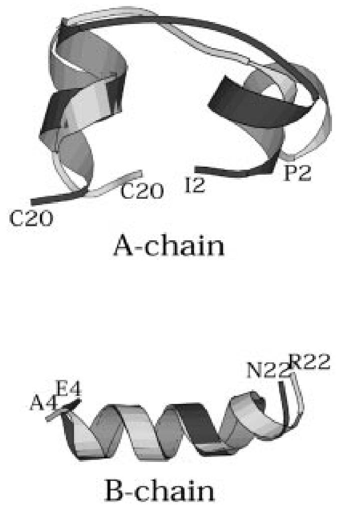

Figure 10. Superpositioning of chain structures from a predicted ROBETTA model for DIGF-3 (yellow) and the hIGF-I crystal structure (blue). DIGF-3 shows the typical features of vertebrate IGFs: A mainly α-helical B-chain and an A-chain with a 'horse shoe' shaped helix-loop-helix motif. N- and C-termini are denoted; N22′ signifies that the residue is positioned in a gapped region of IGF-I in the sequence alignment (Fig. 9A).

CONCLUSION

Contact-based sequence alignment closes a gap between pure sequence and pure structure alignments. Structural information can be used to improve alignments of homologous sequences, even if the structure of only one sequence is known. The described contact-based alignment algorithm is fast and produces improved alignments when compared to Needleman-Wunsch alignments.

The use of side-chain contact information in a sequence alignment routine introduces information about structural constraints between residues, which is ignored in classical PAM-type alignments, where the score of each aligned residue pair is independent of any other residue match. Contrastingly, in CAO contact-based alignment, the score of a residue pair match is dependent on the composition of a second match, because a contact consists of two (correlated) residues and the alignment of two contacts includes four residues. It proved advantageous to use CAO and PAM scores together, although the combination of two different scoring matrices (PAM and CAO) is non-trivial, in particular because the number of added CAO scores to a cell depends on the number of contacts the

respective residue is involved in. This problem was addressed by using a GA.

Sequence alignment often serves as a first step in comparative modelling projects. Ambiguous regions in the resulting alignment are then corrected at a later stage according to the evidence from structural or functional investigations. The flow of information is typically in the order sequence → structure → function. In contact-based alignment, the flow of information is partially reversed to sequence ← structure, because structural information is incorporated *a priori*. The relevance of this method is illustrated here by its application in annotating eight putative IGF sequences from *Drosophila*. The structural information from the template hIGF-I is essential for the correct alignment of canonical cysteines and thus assignment of the A- and B-chain, from which a comparative model was derived. Independently generated *de novo* ROBETTA models support the comparative model.

Alignment routines operate with the information that is passed to them in the form of substitution scores and gap penalties. This fact has several implications: first, the scoring scheme has to match a variety of biological (evolutionary) processes; second, more information from distinct sources should improve alignments; and third, different scoring matrices can be combined but the scoring scheme requires careful parameterization. The alignment routine presented here demonstrates how, in principle, biological or physicochemical information about sequence, structure and function can be summed up by pre-processing in a single DP matrix. Particularly interesting is the fact that long-range correlations between residues can be included. Here we have used contact information, but the scheme can be extended to include other classes of information such as phylogeny, motifs, solvation, fold, function and NMR constraints. One can envision a set of generic properties that characterize the specific features of a protein family. Similar approaches exist for fold recognition and threading tools (34–38), but in terms of sequence alignment, a systematic analysis remains to be performed.

Sequence similarity is a very good indicator for evolutionary relationship, because the enormous dimension of sequence space renders chance similarity highly improbable. Improving the alignment quality of distant sequences therefore allows for a better estimation of the common ancestry of proteins. Moreover, by finding appropriate features that can provide essential information to correct alignments, we might be able to learn more about the features that are important during the course of protein evolution.

Availability

The contact-based alignment program 'ALICAO', the program 'GETCONT' to compute contact maps from coordinate files (Protein Database format) and the CAO matrices are available from our website at http://ibivu.cs.vu.nl/ftp/.

ACKNOWLEDGEMENTS

The authors thank Dr G. Dodson, Dr A. Gould and Dr I. Robinson for stimulating discussions and help. We thank Dr F. Fraternali and V. Simossis for advice.

REFERENCES

1. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
2. Chothia,C. and Lesk,A. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
3. Venclovas,C. (2003) Comparative modeling in CASP5: Progress is evident, but alignment errors remain a significant hindrance. *Proteins Struct. Funct. Genet.*, **53**, 380–388.
4. Vogt,G., Etzold,T. and Argos,P. (1995) An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J. Mol. Biol.*, **249**, 816–831.
5. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
6. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
7. Taylor,W.R. (1986) Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, **188**, 233–258.
8. Barton,G.J. and Sternberg,M.J. (1987) A strategy for the rapid multiple alignment of protein sequences. *J. Mol. Biol.*, **198**, 327–337.
9. Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.
10. Corpet,F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, **16**, 10881–10890.
11. Altschul,S.F., Carroll,R.J. and Lipman,D.J. (1989) Weights for data related by a tree. *J. Mol. Biol.*, **207**, 647–653.
12. Gotoh,O. (1993) Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.*, **9**, 361–370.
13. Gotoh,O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.*, **264**, 823–838.
14. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS*, **10**, 19–29.
15. Heringa,J. (1999) Two strategies for sequence comparison: Profile-preprocessed and secondary structure-induced multiple alignment. *Comput. Chem.*, **23**, 341–364.
16. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate Multiple Sequence Alignment. *J. Mol. Biol.*, **302**, 205–217.
17. Lee,C., Grasso,C. and Sharlow,M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
18. Dayhoff,M.O., Schwart,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5, Ch. 22, National Biomedical Research Foundation, Washington, DC, pp. 345–352.
19. Lin,K.X., Kleinjung,J., Taylor,W.R. and Heringa,J. (2003) Testing homology with CAO: A contact-based Markov model of protein evolution. *Comput. Biol. Chem.*, **27**, 93–102.
20. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
21. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
22. Altschul,S.F., Madden,T.L., Schaeffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Taylor,W.R. (1998) Dynamic sequence databank searching with templates and multiple alignment. *J. Mol. Biol.*, **280**, 375–406.
24. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
25. Brzozowski,A.M., Dodson,E.J., Dodson,G.G., Murshudov,G.N., Verma,C., Turkenburg,J.P., de Bree,F.M. and Dauter,Z. (2002) Structural origins of the functional divergence of human insulin-like growth factor-I and insulin. *Biochemistry*, **41**, 9389–9397.
26. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E.M., Bonneau,R., Rohl,C.A. and Baker,D. (2003) Automated prediction of CASP-5 structures using the Robetta Server. *Proteins*, **53**, 524–533.

27. Rechler,M.M. and Nissley,S.P. (1990) Insulin-like growth factors. In Sporn,M.B. and Roberts,A.B. (eds), *Peptide Growth Factors and Their Receptors*, Vol. I, Springer Verlag, Berlin, pp. 263–367.
28. Smit,A.B., Vreugdenhil,E., Ebberink,R.H., Geraerts,W.P., Klootwijk,J. and Joosse,J. (1988) Growth-controlling molluscan neurons produce the precursor of an insulin-related peptide. *Nature*, **331**, 535–538.
29. Lagueux,M., Lwoff,L., Meister,M., Goltzene,F. and Hoffmann,J.A. (1990) cDNAs from neurosecretory cells of brains of *Locusta migratoria* (Insecta, Orthoptera) encoding a novel member of the superfamily of insulins. *Eur. J. Biochem.*, **187**, 249–254.
30. Kondo,H., Ino,M., Suzuki,A., Ishizaki,H. and Iwami,M. (1996) Multiple gene copies for bombyxin, an insulin-related peptide of the silkmoth *Bombyx mori*: Structural signs for gene rearrangement and duplication responsible for generation of multiple molecular forms of bombyxin. *J. Mol. Biol.*, **259**, 926–937.
31. Pierce,S.B., Costa,M., Wisotzkey,R., Devadhar,S., Homburger,S.A., Buchman,A.R., Ferguson,K.C., Heller,J., Platt,D.M., Pasquinelli,A.A., Liu,L.X., Doberstein,S.K. and Ruvkun,G. (2001) Regulation of DAF-2 receptor signaling by human insulin and *ins-1*, a member of the unusually large and diverse *C. elegans* insulin gene family. *Genes Dev.*, **15**, 672–686.
32. Brogiolo,W., Stocker,H., Ikeya,T., Rintelen,F., Fernandez,R. and Hafen,E. (2001) An evolutionarily conserved function of the *Drosophila* insulin receptor and insulin-like peptides in growth control. *Curr. Biol.*, **11**, 213–221.
33. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
34. Bowie,J.U., Luethy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a common structure. *Science*, **253**, 164–170.
35. Rodionov,M.A. and Johnson,M.S. (1994) Residue–residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds. *Protein Sci.*, **3**, 2366–2377.
36. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
37. Russell,A.J. and Torda,A.E. (2002) Protein sequence threading: averaging over structures. *Proteins Struct. Funct. Genet.*, **47**, 496–505.
38. Teodorescu,O., Galor,T., Pillardy,J. and Elber,R. (2004) Enriching the sequence substitution matrix by structural information. *Proteins*, **54**, 41–48.